RETRIEVAL SYSTEM

Publication number: JP10091638

Publication date: 1998-04-10

Inventor: SUZUOKA SETSU (JP); SUGANO SHINICHI (JP); SAWAJIMA SHINSUKE (JP); YAMANE TETSUYA (JP)

Applicant: TOKYO SHIBAURA ELECTRIC CO (JP)

Classification:

- international: ***G06F12/00; G06F17/30;*** **G06F12/00; G06F17/30;** (IPC1-7): G06F17/30

- European: G06F17/30W1

Application number: JP19960245049 19960917

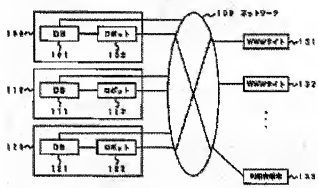
Priority number(s): JP19960245049 19960917

Also published as:

US5933832 (A1)

Abstract of **JP10091638**

PROBLEM TO BE SOLVED: To effectively put data of high frequency of update together in a data base by gathering corresponding data and generating the data base on condition that the update frequencies of the data are within an allocated range of update frequencies. SOLUTION: Retrieval devices 100 to 120 consisting of pairs of robots 102 to 122 and data bases 101 to 121, WWW sides 131 and 132, and a user terminal 133 are connected to a network 100. A range of update frequencies of object page data is assigned to each data base. The robot 102 gathers a site group or data group which change at high frequency and puts it in the data base 101. The robot 122 gathers a site group or data group which changes at low frequency and stores it in the data base 121. The robot 122 gathers a site group or data group which changes at intermediate frequency and stores it in the data base 111.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報(A)**　(11)特許出願公開番号

特開平10−91638

(43)公開日　平成10年(1998)４月10日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３４０Ａ
		15/40	３１０Ｃ
			３８０Ｚ

審査請求　未請求　請求項の数11　ＯＬ　（全 16 頁）

(21)出願番号　特願平8−245049

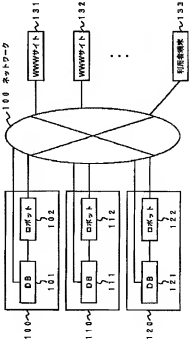
(22)出願日　平成８年(1996)９月17日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地
(72)発明者　鈴岡　節
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(72)発明者　菅野　伸一
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(72)発明者　澤島　信介
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(74)代理人　弁理士　鈴江　武彦　（外６名）

最終頁に続く

(54)【発明の名称】　検索システム

(57)【要約】
【課題】　ネットワーク上に散在する膨大な検索対象データを効率良く取得しデータベース化する検索システムを提供すること。
【解決手段】　ネットワーク上でロボットを用いて収集したデータをもとにデータベースを作成し、検索要求に応じてデータベース検索を行なう検索システムにおいて、データベース化の対象とすべきデータの更新頻度の範囲がデータベース固有に割り当てられ、データの更新頻度が該割り当てられた更新頻度の範囲内にあることまたはデータの属するデータ群における平均的な更新頻度が該割り当てられた更新頻度の範囲内にあることを少なくとも条件として、該当するデータを収集し所定の構造のデータベースを作成するデータベース作成手段を備えたことを特徴とする。



【特許請求の範囲】
【請求項1】ネットワーク上でロボットを用いて収集したデータをもとにデータベースを作成し、検索要求に応じてデータベース検索を行なう検索システムにおいて、
データベース化の対象とすべきデータの更新頻度の範囲がデータベース固有に割り当てられ、データの更新頻度が該割り当てられた更新頻度の範囲内にあることまたはデータの属するデータ群における平均的な更新頻度が該割り当てられた更新頻度の範囲内にあることを少なくとも条件として、該当するデータを収集し所定の構造のデータベースを作成するデータベース作成手段を備えたことを特徴とする検索システム。
【請求項2】既にデータベース化したデータの更新頻度または該データの属するデータ群における平均的な更新頻度が、そのデータベースに割り当てられた前記更新頻度の範囲外のものとなった場合には、該データを対象とし得る他のデータベースにて該データをデータベース化の対象とさせるための処理を行う処理手段をさらに備えたことを特徴とする請求項1に記載の検索システム。
【請求項3】利用者から与えられた検索要求に応答して、互いに同一でない前記更新頻度の範囲が割り当てられて作成された複数の前記データベースを連携させて検索を行い、得られた検索結果を返す検索手段をさらに備えたことを特徴とする請求項1に記載の検索システム。
【請求項4】前記検索要求で更新頻度範囲および更新時刻範囲の少なくとも一方が指定されている場合には、前記検索手段は、指定された更新頻度範囲および更新時刻範囲の少なくとも一方に該当するデータについてのみ検索を行い、
前記検索要求で検索範囲の指定がない場合には、前記検索手段は、全データを対象として、または更新頻度範囲および更新時刻範囲の少なくとも一方のデフォルト値で制限された範囲を対象として検索を行うことを特徴とする請求項1に記載の検索システム。
【請求項5】前記検索システムを構成するハードウェアのうち更新頻度の高いデータに対応する部分ほど、高い処理能力を持たせることを特徴とする請求項1ないし4のいずれか1項に記載の検索システム。
【請求項6】前記高い処理能力は、より高速な計算機を用いることおよびより多数の計算機を用いることの少なくとも一方によって実現することを特徴とする請求項5に記載の検索システム。
【請求項7】ネットワーク上でロボットを用いて収集したデータをもとにデータベースを作成し、データベース検索を行なう検索システムにおいて、
外部からの参照要求に応答して取得されたデータおよびロボットを用いて収集されたデータを保持するキャッシュ手段と、
外部から参照要求が与えられた場合に、前記キャッシュ手段に該当するデータが保持されているならば、前記キャッシュ手段からデータを提供し、前記キャッシュ手段に該当するデータが保持されていないならば、該データを保持する本来のサーバーから該データを取得して提供するデータ提供手段とを備えたことを特徴とする検索システム。
【請求項8】外部から参照要求されたデータについての統計処理を行って、今後参照要求されるデータを予測する予測手段と、
予測されたデータおよび予め明示的に指定されたデータを、ロボットを用いて取得し前記キャッシュ手段にプリフェッチするプリフェッチ手段とをさらに備えたことを特徴とする請求項7に記載の検索システム。
【請求項9】前記プリフェッチ手段は、取得対象となるデータの更新頻度に応じた頻度で該データを取り直すことを特徴とする請求項7に記載の検索システム。
【請求項10】前記検索要求に応答して行う検索で対象とするデータの範囲の制約条件として、ロボットで収集されたデータに限る条件、外部からの参照要求に応答して取得されたデータに限る条件、同じ名前またはアドレスを持つデータについては最新のものだけに限る条件、動的または対話的に生成されたデータ以外のものに限る条件、および指定されたサイト群またはデータ群に限る条件のうち少なくとも1つを課すことを特徴とする請求項7に記載の検索システム。
【請求項11】前記キャッシュ手段は、取得されたデータにその更新時刻情報および収集時刻情報の少なくとも一方を付加して保持することを特徴とする請求項7に記載の検索システム。
【発明の詳細な説明】
【0001】
【発明の属する技術分野】本発明は、ネットワーク上に分散したデータの検索システムに関する。
【0002】
【従来の技術】Ａｌｔａｖｉｓｔａ（ｈｔｔｐ：／／ｗｗｗ．ａｌｔａｖｉｓｔａ．ｃｏｍ／）、Ｌｙｃｏｓ（ｈｔｔｐ：／／ｗｗｗ．ｌｙｃｏｓ．ｃｏｍ／）、Ｙａｈｏｏ！（ｈｔｔｐ：／／ｗｗｗ．ｙａｈｏｏ．ｃｏｍ／）などロボットを用いたネットワーク上の検索システムは多数存在する。これらはロボットと呼ばれる機械的にネットワーク上で情報を収集するソフトウェアを用いている。そして、収集したデータをデータベース化し、利用者が検索できるようにしている。
【0003】上記ロボットは、ネットワーク上でＨＴＭＬ（Ｈｙｐｅｒ　Ｔｅｘｔ　Ｍａｒｋｕｐ　Ｌａｎｇｕａｇｅ）で記述された文章を探し、そこに記載されているリンク先を辿って、ネットワーク上に存在するデータを収集する。データベース化については、フルテキストサーチをするものもあれば、タイトルやＵＲＬといった部分のみを検索対象とするようなものもある。
【0004】上記データベースは、量が多いので分散化

されている場合もある。しかし、あくまでも量が多いための単なる分割であり、何らかの意味を持って分割してはいない。

【0005】上記検索には、キーワード検索が行なわれる。すなわち、探したい文章に含まれているであろう語を入力して、検索を行なう。一方、人気のあるサイトへのアクセス集中を分散させ、トラフィックを軽減するために、ミラーサイトが設けられることがある。例えば、Ｐｏｉｎｔ Ｃａｓｔ Ｎｅｔｗｏｒｋ（ＰＣＮ）社のＩ－Ｓｅｒｖｅｒ（ｈｔｔｐ：／／ｗｗｗ．ｐｏｉｎｔｃａｓｔ．ｃｏｍ／ｐｒｏｄｕｃｔｓ／ｉｓｅｒｖｅｒ．ｈｔｍｌ）ではＰＣＮ本社へ定期的に情報をプリフェッチして、ミラーサイトを管理している。

【0006】

【発明が解決しようとする課題】従来、ネットワーク上に分散したデータの検索システムにおいては、以下のような問題点があった。

（１）増大するデータを扱うのが困難になりつつある。例えばＷＷＷ上のページデータが１９９６年で世界で４０００万以上あると言われ、今後も指数関数的に増加すると予想される。現在、ページ数も、１ページあたりのデータ量も急激に増大する傾向にある。このように急増するデータを単に量により分割するだけでは、データベース管理が極めて困難である。

【0007】（２）更新頻度が高い情報を扱うのが困難である。一日に何度も更新されるデータについては、現在の検索システムではロボット探索対象から外している。この理由は、頻繁に更新されるデータをロボットで情報収集してデータベース化しても、そのデータが検索され前に更新されることが少なくないからである。このような場合には、検索結果に現れたページを見ても、既になくなっていたり、内容が全く別のものに変更されたために利用者の意図したものとは別ものが表示されたりする不都合が生じる。

【0008】本発明は、上記事情を考慮してなされたもので、ネットワーク上に散在する膨大な検索対象データを効率良く取得しデータベース化する検索システムを提供することを目的とする。また、本発明は、極めて更新頻度の高いデータをも効果的にデータベース化する検索システムを提供することを目的とする。

【0009】

【課題を解決するための手段】本発明（請求項１）は、ネットワーク（例えば、インターネットのＷＷＷ）上でロボットを用いて収集したデータ（例えばページのようなハイパーメディアデータ）をもとにデータベースを作成し、検索要求に応じてデータベース検索を行なう検索システムにおいて、データベース化の対象とすべきデータの更新頻度（例えば、統計的な更新頻度、あるいは最終更新時刻）の範囲がデータベース固有に割り当てられ、データの更新頻度が該割り当てられた更新頻度の範囲内にあることまたはデータの属するデータ群（例えば、サイト）における平均的な更新頻度が該割り当てられた更新頻度の範囲内にあることを少なくとも条件として、該当するデータを収集し所定の構造のデータベースを作成するデータベース作成手段を備えたことを特徴とする。

【0010】データベースは、例えば、データのアドレスとキーワードの組からなる構造を持つ。具体的には、例えば、ページのＵＲＬにキーワードを付加したものである。

【0011】本発明（請求項２）は、請求項１に記載の検索システムにおいて、既にデータベース化したデータの更新頻度または該データの属するデータ群における平均的な更新頻度が、そのデータベースに割り当てられた前記更新頻度の範囲外のものとなった場合には、該データを対象とし得る他のデータベースにて該データをデータベース化の対象とさせるための処理を行う処理手段をさらに備えたことを特徴とする。

【0012】本発明（請求項３）は、請求項１に記載の検索システムにおいて、利用者から与えられた検索要求に応答して、互いに同一でない前記更新頻度の範囲が割り当てられて作成された複数の前記データベースを連携させて検索を行い、得られた検索結果を返す検索手段をさらに備えたことを特徴とする。

【0013】本発明（請求項４）は、請求項１に記載の検索システムにおいて、前記検索要求で更新頻度範囲および更新時刻範囲の少なくとも一方が指定されている場合には、前記検索手段は、指定された更新頻度範囲および更新時刻範囲の少なくとも一方に該当するデータについてのみ検索を行い、前記検索要求で検索範囲の指定がない場合には、前記検索手段は、全データを対象として、または更新頻度範囲および更新時刻範囲の少なくとも一方のデフォルト値で制限された範囲を対象として検索を行うことを特徴とする。

【0014】本発明（請求項５）は、請求項１ないし４のいずれか１項に記載の検索システムにおいて、前記検索システムを構成するハードウェアのうち更新頻度の高い（例えば、統計的な更新頻度の高い、あるいは最終更新時刻の新しい）データに対応する部分ほど、高い処理能力を持たせることを特徴とする。

【0015】本発明（請求項６）は、請求項５に記載の検索システムにおいて、前記高い処理能力は、より高速な計算機を用いることおよびより多数の計算機を用いることの少なくとも一方によって実現することを特徴とする。

【0016】本発明によれば、データの更新頻度の高さ（あるいは最終更新時刻の新しさなど）を人気の度合いと見做せば、人気の度合い応じて異なったデータベースにてデータを管理することができる。また、人気の高いすなわちアクセス頻度の高いデータベースを処理する計

算機を強力にし、アクセス頻度の低い膨大な量のデータについては処理能力の低い計算機を割り当てることにより、ハードウェア資源を効率的に使用したシステムを構築できる。これによって、膨大なデータを対象とした効果的な検索システムを提供することができる。

【００１７】本発明（請求項７）は、ネットワーク上でロボットを用いて収集したデータをもとにデータベースを作成し、データベース検索を行なう検索システムにおいて、外部からの参照要求に応答して取得されたデータおよびロボットを用いて収集されたデータを保持するキャッシュ手段と、外部から参照要求が与えられた場合に、前記キャッシュ手段に該当するデータが保持されているならば、前記キャッシュ手段からデータを提供し、前記キャッシュ手段に該当するデータが保持されていないならば、該データを保持する本来のサーバーから該データを取得して提供するデータ提供手段とを備えたことを特徴とする。

【００１８】本検索システムは、プロキシーも兼ねるものであり、これによって、利用者が要求したデータがシステム内にあるならば、それが利用者からの要求によって取得したものであっても、それがロボットによって収集されたものであっても、それを利用者に提示することができる。

【００１９】これによって、極めて更新頻度が高いデータに対しても、検索を適用することができる。本発明（請求項８）は、請求項７に記載の検索システムにおいて、外部から参照要求されたデータについての統計処理を行って、今後参照要求されるデータを予測する予測手段と、予測されたデータおよび予め明示的に指定されたデータを、ロボットを用いて取得し前記キャッシュ手段にプリフェッチするプリフェッチ手段とをさらに備えたことを特徴とする。

【００２０】本発明では、取得可能なすべてのデータをロボットを用いてあらかじめ収集せずに、あらかじめ指定したデータおよび利用者からの統計的観点から参照要求があると思われるデータについてロボットによりデータをプリフェッチしておくので、適切なデータに対して効果的にミラー化される。

【００２１】本発明（請求項９）は、請求項７に記載の検索システムにおいて、前記プリフェッチ手段は、取得対象となるデータの更新頻度に応じた頻度で該データを取り直すことを特徴とする。

【００２２】本発明（請求項１０）は、請求項７に記載の検索システムにおいて、前記検索要求に応答して行う検索で対象とするデータの範囲の制約条件として、ロボットで収集されたデータに限る条件、外部からの参照要求に応答して取得されたデータに限る条件、同じ名前またはアドレスを持つデータについては最新のものだけに限る条件、動的または対話的に生成されたデータ以外のものに限る条件、および指定されたサイト群またはデータ群に限る条件のうち少なくとも１つを課すことを特徴とする。

【００２３】本発明（請求項１１）は、請求項７に記載の検索システムにおいて、前記キャッシュ手段は、取得されたデータにその更新時刻情報および収集時刻情報の少なくとも一方を付加して保持することを特徴とする。

【００２４】これによって、取得元のデータの名前が同じでも時刻によって異なるデータに対しても管理できる。なお、以上の各装置に係る発明は、方法に係る説明としても成立する。また、上記の発明は、相当する手順あるいは手段をコンピュータに実行させるためのプログラムを記録した機械読取り可能な媒体としても成立する。

【００２５】

【発明の実施の形態】以下、図面を参照しながら発明の実施の形態を説明する。まず、語句の定義を行う。プロキシー（Ｐｒｏｘｙ）とは、クライアント（例えば利用者端末）からサーバー（例えばWWWサイト）への資源アクセスの際にアプリケーションレベルにおいて、クライアントとサーバーの間に入り、クライアントからの資源アクセス要求をサーバーに対して中継し、サーバーからの応答をクライアントに対して中継する機能を有するサーバーのことを言う。

【００２６】ページ（ｐａｇｅ）とは、ハイパーテキストのページを意味するものとする。WWWの世界では、１つのページはユニークなURLを持つ。URL（Ｕｎｉｆｏｒｍ Ｒｅｓｏｕｃｅ Ｌｏｃａｔｉｏｎ）とは、ページデータをアクセスするのに必要な情報である。URLは、プロトコル、ドメイン名、ポート番号、パス名の情報を含む。

【００２７】CGI（Ｃｏｍｍｏｎ Ｇａｔｅｗａｙ Ｉｎｔｅｒｆａｃｅ）とは、対話的なページや動的なページを作るためにサーバーからプログラムを起こすためのインターフェースである。

【００２８】ロボット（Ｒｏｂｏｔ）とは、Ｈｙｐｅｒ Ｔｅｘｔ Ｍａｒｋｕｐ Ｌａｎｇｕａｇｅ（ＨＴＭＬ）やＳｔａｎｄａｒｄ Ｇｅｎｅｒａｌｉｚｅｄ Ｍａｒｋｕｐ Ｌａｎｇｕａｇｅ（ＳＧＭＬ）のようなハイパーテキストで記述された文書を読み、そこに書かれているリンクを機械的に辿りながら文書をネットワーク上で収集するものであり、ソフトウェアにより実現される。ロボットの代わりにスパイダー（ｓｐｉｄｅｒ）あるいはワンダラー（Ｗａｎｄｅｒｅｒ）などと呼ばれることもある。

【００２９】ロボットの基本的な動作は次のようになる。

（手順１）指定されたURLの根を探訪リストに登録する。

（手順２）ロボットは、探訪リストに従いページを取得する。

（手順３）取得されたページを解析してＵＲＬを抽出する。
（手順４）抽出されたＵＲＬを探訪リストに追加する（ただし、ＵＲＬの重複登録はしない）。
以降、手順２～４を繰り返す。なお、ページの取得頻度は、該ページの更新頻度に応じて決めるようにしても良い。

【００３０】次に、本実施形態を概略的に説明する。本実施形態では、ネットワーク中に分散されたデータの一例としてページを扱うものとする。

【００３１】前述したように、例えば、Ｗｏｒｌｄ Ｗｉｄｅ Ｗｅｂ（ＷＷＷ）上のページ数（ページの種類）は４０００万を越えると言われる。この数は、今後も指数関数的に増え続けると予測されている。このような膨大な量のページを単一のデータベースで管理することは極めて困難である。

【００３２】データベースを分割する最も単純な方法は、サイト（ドメイン）単位でデータベースを分割することであるが、こうすると、どのデータベースも等しく高速でなければならない。データベースを分割することができても、すべてが高速でなければならないとすると、データベース構築の負担は依然高い。

【００３３】そこで、第１の実施形態では、データベースの内容を人気の度合いに応じて分割するようにしている。そして、人気の高いデータベースは高速なシステム（例えば大容量メモリを持つマシン）の上に載せ、人気があまりないデータベースは低速なシステムの上に載せるようにする。このようにすると、人気の高いデータベースを載せるマシンだけ高速なマシンを使えば良くなり、データベース構築の負担を効果的に軽減することができる。

【００３４】ここで、ページの人気の高さを知るためには厳密に言うとネットワークの視聴率調査などをしなければならないが、そのような作業は大きな困難を伴い現実的ではない。そこで、本実施形態では、次のような良く成り立つ近似を使う。まず「ページが読まられずに高い人気を保ためには、絶えずコンテンツをアップデートしていく必要がある」と考える。そして、その逆をとって「データの更新頻度が高いページは、人気の高いページである」と近似する。つまり、本実施形態では、人気のバロメーターとしてデータの更新頻度を使い、データベースの内容をデータの更新頻度に応じて分割する。なお、ページの更新頻度はロボットを走行させることにより取得できる情報である。

【００３５】ところで、更新頻度が高いページには１日に何度も更新されるものもある。このようなページに対して時々しかアクセスしない方法を採る場合、実際のページデータと検索システム内のデータベースとが不一致となる状態が発生する。特に、データベース検索の結果をもとにページを参照しにいくと、既に該当ページがなくなっていたり、ページ自体はあっても内容が別のものに変更されてたりすることがあり、このような場合に不具合が発生する。

【００３６】一方、データベースの陳腐化による矛盾を軽減するためには、ロボットが非常に高頻度にページをアクセスする必要がある。しかし、不定期に頻繁に変更されるページの最新情報に追い付くために頻繁にアクセスすることは、無用なトラフィックを増大させ、情報を保持するサイトにも検索システム側にも不利益を被らせる。

【００３７】そこで、第２の実施形態では、データベース化した元データを保存しておき、それを利用者に提示するようにしている。このようにすると、実際のページの変化には多少遅れるが、無駄にトラフィックを増やすこともなく、しかも検索結果に対応した元ページを常に見ることができる。

【００３８】なお、第１の実施形態と第２の実施形態を組み合わせることも可能である。この場合には、両者の効果を得ることができる。以下、本発明の実施形態について詳しく説明する。

【００３９】（第１の実施形態）まず、第１の実施形態について説明する。本実施形態のシステム構成例を、図１、図４、図６に示す。

【００４０】本実施形態では、複数のデータベースを容易し、データの更新頻度に応じてデータベースを使い分ける。すなわち、各データベースに、対象とするページデータの更新頻度の範囲を割り当てる。そして、ユーザが要求するキーワードについて検索を行なう際には、複数のデータベースを連携させて検索し、結果をまとめて利用者に提示する。

【００４１】各データベースへのページ分担方法には、例えば次のようなものが考えられる。
（ａ）統計的更新頻度情報によって分担
（ｂ）最終更新時刻によって分担
（ｃ）統計的更新頻度情報と最終更新時刻との総合的情報によって分担ここで、（ｂ）の最終更新時刻によって分担する方法について説明する。

【００４２】あるページは、更新された直後は頻繁にアクセスされ（つまり人気があり）、最後に更新されてから時間が経過している程、アクセスされる頻度が少ない（つまり人気がない）と考えられる。そこで、例えば図３のように、最終更新時刻の範囲に応じて、格納すべきデータベースを分担する。

【００４３】あるページに関する情報を格納するデータベースを決定する方法には、例えば次のようなものが考えられる。
（１）サイト単位に格納すべきデータベースを決定する。この場合には、サイト内のデータの更新頻度の平均値を評価値に用いる。
（２）サイト内のディレクトリ単位に格納すべきデータ

ベースを決定する。この場合には、ディレクトリ内のデータの更新頻度の平均値を評価値に用いる。

（3）データ単位に格納すべきデータベースを決定する。この場合には、そのデータの更新頻度を評価値に用いる。

【0044】ここで、更新頻度は、上記の統計的更新頻度情報や最終更新時刻などである。なお、上記の（1）～（3）の方法は、併用可能である。例えば、サイトAについてはサイト単位にデータベースに入れ、サイトBについては、データ単位にデータベースに入れるようにしても良い。また、サイトC内で、ディレクトリaについてはディレクトリ単位にデータベースに入れ、ディレクトリbについてはデータ単位にデータベースに入れるようにすることも可能である。

【0045】また、更新頻度が高いデータほど、内部ネットワークにつながれたサーバにおくことも考えられる。例えば、更新頻度が高い方のデータを組織内のイントラネットにおき、更新頻度が低い方のデータをインターネットに直接接続された場所で管理する。

【0046】なお、本実施形態では、データベースにはページ自体ではなくキーワードとURLとを格納するものとする。また、ページを全文検索などして抽出したキーワードをURLに付加して格納し、キーワードでURLを検索するものとする。

【0047】また、本実施形態では、語単位もしくはキーワード単位のデータベースについて述べているが、文字単位のデータベースであっても良い。次に、図1、図4、図6に示す各システム構成例について説明する。

【0048】図1の構成例では、ネットワーク100に、複数のロボットとデータベースとの組（101と102，111と112，121と122）からなる検索装置100，110，120、複数のWWWサイト（131，132）、利用者端末（133）が接続されている。

【0049】各データベースには、前述したようなページ分担方法で、対象とする更新頻度を割り当てる。第1のロボット102は、高頻度に変化するサイト群もしくはデータ群を集め（例えばWWWサイト131，132から集め）、それをデータベース化して第1のデータベース101に格納する。

【0050】第3のロボット122は、低頻度に変化するサイト群もしくはデータ群を集め、それをデータベース化して第3のデータベース121に格納する。第2のロボット112は、それ以外の中頻度に変化するサイト群もしくはデータ群を集め、それをデータベース化して第2のデータベース111に格納する。

【0051】高頻度、低頻度、それ以外の中頻度に夫々対応する実際の統計的更新頻度情報（あるいは、最終更新時刻など）の範囲は、適宜設定する。次に、動的なデータベースの分担変更について述べている。

【0052】本実施形態では、統計から得られる更新頻度情報に応じて分割された各データベースに該当するページのURLを入れるが、時間とともにページの更新頻度（あるいはページの属するサイトの平均的な更新頻度等）は変化することがあるので、あるページの更新頻度（あるいはページの属するサイトの平均的な更新頻度等）がそのページを分担した初期のデータベースの持つ更新頻度の範囲を逸脱する場合が発生する。従って、あるページを分担中のデータベースから適切な更新頻度範囲を持つデータベースにそのページデータもしくはサイトを受け持つように依頼するようにするのが望ましい。この依頼は、データベース間の交渉により実現されるものとする。

【0053】例えば、図1において、第1のロボット102は、統計的に高頻度のデータ群を取り寄せ第1のデータベース101に格納する。しかし、当初高頻度で更新されていたデータの更新頻度が自分が受け持つ範囲よりも低下したならば、そのデータを第2のロボット112とデータベース111に引き受けてもらう。また、更新頻度が大きく落ちた場合には、第3のロボット122とデータベース121に担当を替えるよう依頼する。

【0054】図2に、図1のように更新頻度に応じてロボットが複数台あり、それぞれにデータベースがある場合の各検索装置の処理手順の一例を示す。ステップS21で、他の検索装置からページの分担を依頼されているかどうか調べ、あればステップS27を行い、なければステップS22を行う。

【0055】ステップS22で、それぞれのロボットは、指定されたページを1つ選び、そのページを取得する。このときのページの統計的更新頻度に比例した頻度でページを取得するようにスケジュールする。なお、そのページについて統計的更新頻度の情報がない場合には、そのページを含むサイトのページのうち得られている統計的更新頻度の平均的な値あるいはデフォルト値などで代用すれば良い。

【0056】ステップS23で、取得したページが前回と変わっているか否かにより、そのページの統計的更新情報を更新する。もし、ネットワークや相手サーバのトラブルにより、そのページの取得に失敗した場合には、そのページの取得に失敗したという記録を残して、ステップS22に戻る。

【0057】ステップS24で、新しい更新頻度が自らが担当している範囲内かどうかを調べる。ステップS25で、もし自らの担当範囲外になったならば、それを範囲内に含む検索装置に以降の処理を依頼する。このとき、そのページのデータは消去する。

【0058】ステップS26で、もし自らの担当範囲内ならば、取得したページをデータベース化し、格納する。例えば、ページデータを形態素解析し、単語レベルに分解し、単語を含むページという形にデータベース化

する。このとき、そのページの前のデータは消去する。

【００５９】ステップＳ２７で、他の検索装置から依頼があった場合には、そのページを自ロボットで扱うことができるように、そのページを登録し、そのページの統計的更新頻度情報を設定する。

【００６０】本実施形態において、検索利用者がデータベース検索を行う場合、利用者端末１３３から複数のデータベース１０１，１１１，１２１のすべてに検索要求を出す方法と、いずれか１つのデータベース１に検索要求を出す方法が考えられる。後者のいずれか１つのデータベースに検索要求を出す場合には、その検索要求を受け取ったデータベースのみが結果を返すようなモードと、そのデータベースが他のデータベースにも問い合わせに行き結果をマージして返すようなモードが考えられる。

【００６１】次に、図４の構成例について説明する。図４は、基本的には図１と同様であり、データの更新頻度に応じた複数のデータベース２０１～２０３が用意されているが、ロボット２０４を一台で兼用する点に関して図１の構成例と相違する。

【００６２】図５に、図４のように、ロボットが１台でデータベースが複数ある場合の検索装置の処理手順の一例を示す。ステップＳ１１で、指定されたページを１つ選び、ロボット２０４を用いてそのページを取得する。このときのページの統計的更新頻度に比例した頻度でページを取得するようにスケジュールする。なお、そのページについて統計的更新頻度の情報がない場合には、そのページを含むサイトのページのうち得られている統計的更新頻度の平均的な値あるいはデフォルト値などで代用すれば良い。

【００６３】ステップＳ１２で、取得したページが前回と変わっているか否かにより、そのページの統計的更新情報を更新する。もし、ネットワークや相手サーバのトラブルにより、そのページの取得に失敗した場合には、そのページの取得に失敗したという記録を残して、ステップＳ１１に戻る。

【００６４】ステップＳ１３で、ステップＳ１１で取得したページの新しい統計的更新確率により、そのページをどのデータベースに担当させるかを決定する。ステップＳ１４で、ページ情報をデータベース化する。例えば、ページデータを形態素解析し、単語レベルに分解し、単語を含むページという形にデータベース化する。このデータをステップＳ１３で決めたデータベースに格納する。このとき、そのページの前のデータは消去する。もし、ここで、これまで格納されていたデータベースと異なるデータベースに格納されていたならば、それをも消去する。もし、取得したページが前回から変更がない場合には、データベース化は行わないが、格納すべきデータベースがそれにより変更された場合には、データの移動のみを行う。

【００６５】以上のように、ロボットの数はデータベースの数と一致している必要はない。例えば、図４の場合、ロボットの数は２台でも４台以上でも良い。各ロボットとデータベースとの対応関係は適宜設定すれば良い。

【００６６】なお、検索利用者によるデータベース検索については前述した図１と同様である。次に、図６の構成例について説明する。図６の検索装置３００は、データベース全体を取りまとめるデータベース・フロントエンド（ＤＢＦ）３０１が設けられている点が図４の検索装置２００と相違する。

【００６７】本構成例では、このＤＢＦ３０１が利用者端末１３３からの検索要求を受付け、適切なデータベースに問い合わせて、結果を利用者に提示する。次に、データベース検索における検索対象範囲の指定について説明する。

【００６８】本第１の実施形態では、検索要求にて、キーワードを用いた検索条件の他に、対象とする更新頻度の範囲および／または更新時刻の範囲を指定できるようにすると好ましい。また、検索要求において明示的に更新頻度が指定されていない場合に、データベースあるいはＤＢＦの方でデフォルト値（例えば最も更新頻度の高いデータベースのみといった更新頻度範囲）をもって検索を行なうようにしても良い。

【００６９】ここで、図７に、図６の検索装置における検索手順の一例を示す。利用者が利用者端末１３３からデータベース・フロントエンド３０１に向けて検索要求を送り出すと、ステップＳ３１で、データベースフロントエンド３０１は利用者端末３０８からの検索要求を受け取る。

【００７０】ステップＳ３２で、その検索要求が更新頻度範囲指定を持つかどうかを判定する。もし持つならば、ステップＳ３３で、利用者の検索要求の対象範囲に応じて適切な範囲のデータベースでのみ検索を行う。

【００７１】もし持たないならば、ステップＳ３４で、すべてのデータベースで検索を行う。ステップＳ３５で、結果をマージして利用者端末３０８に返す。

【００７２】次に、システムのハードウェア構成に関して説明する。本第１の実施形態では、更新頻度の高い方（例えば、統計的更新頻度情報の高い方、あるいは最終更新時刻の新しい方など）を受け持つデータベース（またはデータベースおよびロボット）などを構成する計算機には、更新頻度の低い方（例えば、統計的更新頻度情報の低い方、あるいは最終更新時刻の古い方など）を受け持つデータベース（またはデータベースおよびロボット）などを構成する計算機よりも、高速性について同等以上のものを用い、あるいは台数について同数以上を用いるなどして、更新頻度が高いデータを検索するデータベースを担当する計算機の方がそうでないデータベースを担当する計算機よりも処理能力が同じかより高いよう

にシステムを構成すると好ましい。

【0073】すなわち、更新頻度が高い方のデータを担当するデータベースの方が更新頻度が低い方のデータを担当するデータベースよりも頻繁に利用されるので、更新頻度が高い方のデータを担当するデータベースの方のみについて処理能力を上げるだけで、全体の処理能力を効果的に向上させることができる。

【0074】従って、本実施形態のように更新頻度に応じてデータベースを分割することにより、更新頻度の高いデータベースを載せる計算機だけ高速なものを使えば良くなり、データベース構築の負担を効果的に軽減することができる。

【0075】例えば、図8のように、第1の検索装置410を構成する計算機群が更新頻度が高いデータ群を担当し、第2の検索装置401を構成する計算機群が更新頻度が低いデータ群を担当している場合には、第1の計算機群410においてはデータベースをハードウェア的に二重化して高速化している。高速化の手段としては、ハードウェアを多重化する他にも、速い素子を使った計算機を使うとか、メモリの容量を大きくするなどの方法がある。

【0076】以上では、本実施形態についてネットワークを1つとして説明したが、図9のように複数のネットワーク500～504が結合された環境であっても良い。さらに、ネットワーク500～504が組織や国のように物理的にまったく離れた場所を結合しているものであっても良い。

【0077】（第2の実施形態）次に、第2の実施形態について説明する。本実施形態では、検索システムにアプロキシー機能も装備し、検索結果として参照されるべきページデータを既に持っているならば、そのデータをネットワークを介して新たに取りに行くことはせずに、既に持っているデータを返す。

【0078】これにより、前述した頻繁に変化するページの問題にも対処することができる。すなわち、頻繁に変化するページでは、検索結果として示されるリンクを辿ったときには、既にそのページがなくなっていたり、更新されていて役にたたないことがある。これに対して、検索用データベースで用いたデータを提示するのであれば、このような問題は生じない。

【0079】すなわち、頻繁に変化するページは、図13に示すようにサンプリング的に取得し、次の取得まで内容を保持しておく。これにより、例えば図13中のもう1でページが消失しあるいは内容が別のものに移行されるなどしても、最後にサンプリングしたもう0のときの内容を提示することができる。

【0080】図10に、本実施形態のシステム構成例を示す。図10に示すように、本実施形態の検索装置601は、ネットワーク600に接続されており、ロボット602、キャッシュ603、データベース化部604、データベース605、データベース・フロントエンド（DBF）607、WWWフロントエンド606を有する。また、図10には示していないが、ネットワーク600を介して各WWWサイトや利用者端末が接続されているものとする。また、図10中では、データベースを1つとして表わしているが、複数に分割されていても良い。また、複数のデータベースに第1の実施形態にて説明した発明を適用し、データの更新頻度に応じてデータベースに情報の格納を分担させても良い。

【0081】本実施形態では、データベースにはページのURLを格納するものとする。また、ページを全文検索などして抽出したキーワードをURLに付加して格納し、キーワードでURLを検索するのもとする。

【0082】最初にデータベース化までを説明し、次に利用方法について説明する。データベース化まで手順の一例を以下に示す。まず、ロボット602を用いて、探訪リストに従って、ネットワーク600を介して他のWWWサイトからデータを収集する。もし自身も独自コンテンツを持つWWWサイトであるならば、自身からもデータを収集する。その収集したものをキャッシュ603に格納する。キャッシュ603に格納されているものの中からデータベース化部604により検索用データベース605を作成する。例えば、語単位でのキーワード検索を行なう場合には、データベース化部604では、キャッシュ603内のデータを形態素解析し、語単位でデータベース化する。これにより、利用者から特定の語を含む情報を要求された場合に、即座にデータベース検索が可能となる。ここで、本検索装置では、データベース化するときのデータの在処として、そのデータを取得したネットワーク上のアドレス（URL）ではなく、キャッシュ603に格納されているデータのアドレスを用いる。

【0083】一方、ユーザからの参照要求によりWWWフロントエンド606がアクセスして取得したページも、キャッシュ603に格納するとともに、上記と同様にデータベース化しておく。

【0084】次に利用する際の手順の一例を以下に示す。利用者は、ネットワーク600を介して、検索装置601のWWWフロントエンド606にアクセスし、検索要求を出す。その要求は、データベース・フロントエンド（DBF）607に伝えられ、複数のデータベースがある場合には、適切なデータベースが選択され、それに検索要求を出す。データベース・フロントエンド（DBF）607では、複数のデータベースに検索要求を出した場合には、それらの結果を取りまとめて、WWWフロントエンド606を介して利用者に検索結果を提示する。利用者は、検索結果の中で、さらにその中身を見てみたいと思うものがあれば、検索装置601のWWWフロントエンド606に参照要求を出す。WWWフロントエンド606では、参照を要求されたページが自キャッ

シュ603に格納されているものであるならば、該ページをキャッシュ603から取り出して参照要求者に返す。もし自キャッシュ603になければ、その旨を参照要求者に返す。

【0085】ここで、検索装置では、取得可能なすべてのデータをロボットを用いて収集せずに、予め指定されたデータに加えて、統計的観点から参照要求があると思われるデータについてロボットによりデータをプリフェッチしておくようにしても良い。これは、WWW上のすべてのデータを検索対象としない場合や、実際のページの更新頻度ではなく、利用者の要求に基づいてデータを更新する場合に有効である。

【0086】すなわち、WWW上のすべてのデータを検索対象としない場合には、どの範囲をロボットで収集するかが問題となる。そこで、この検索サーバ兼プロキシーへの要求に現れるページやサイトを統計処理し、その頻度が高いデータやサイトのデータを優先的にロボットを用いてあらかじめプリフェッチしておく。このときには、実際のページの更新情報が高いもの程よくそのページをロボットが訪問するのみならず、そのページに対する参照要求の発生確率が高いページほど良くそのページをロボットが訪問するようにする。これにより、システム管理者が特別に指定しなくても、適切なデータに対してミラー化される。

【0087】上記のような検索装置に構成例を図11に示す。図11の検索装置701は、図10の検索装置601にユーザ要求記録部708を追加したものである。従って、相当する部分の説明は省略し、相違する部分を中心に説明を行う。

【0088】図12に、本検索装置701による情報収集の処理手順を示す。ステップS41で、利用者のアクセスログを解析し、そのサイトで良く参照されるページやサイトの情報を得る。

【0089】ステップS42で、上記とは別にシステム管理者などにより明示的に指示されたページやサイトの情報をステップS41で得たものとマージする。ステップS43で、上記で得たデータを、その統計的更新確率にしたがってロボットを用いて取得する。もし、ページについて統計的更新確率情報が得られていなかったときには、そのページを含むサイトのページのわかっている統計的更新確率情報の平均値で代用する。さらに、そのサイトの統計的更新確率情報もわからない場合には、知っているすべてのサイトの統計的更新確率情報もしくはデフォルト値で代用する。この統計的更新確率情報に比例した頻度でデータを繰り返し取得する。また、あるサイトがある時刻に更新される可能性が高いことがわかったならば、その時刻よりも少し後に情報を取に行くようにする。

【0090】さて、本検索装置701は、プロキシーも兼ねているので、利用者は検索要求でなく、単にネットワーク上の情報が欲しいときには、参照要求を検索装置701に出す。その参照要求は、WWWフロントエンド706を介して、ユーザ要求記録部708に出され、ここで要求データの記録が残される。ここで要求されたデータがキャッシュ703にあれば、それをそのまま返し、なければネットワーク700を介してデータを取りに行き、そのデータをキャッシュ703に一旦格納した後、WWWフロントエンド707を介して利用者に返す。

【0091】このように、図11の検索装置では、利用者がどのデータに関心が高いかといった情報がユーザ要求記録部708に格納されている。従って、ロボットでデータを予め収集するときに、ロボットで取得できるすべてのデータを取ろうとするのではなく、ユーザ要求記録部708に格納されているデータと明示的に指示された取得すべきデータとを取得する。

【0092】なお、取得すべきでないデータ群を指定して、それらはユーザ要求記録部708にあるものであっても取得しないようにしても良い。ところで、頻繁に更新されるデータについては、ユーザ要求記録部708の記録を見ても有効でないと考えられる。なぜならば、再び訪れたときにはそのデータが消滅している可能性が高い。従って、そのようなデータについては、サイトもしくはデータへのパスのみを有効な情報とし、同じデータでなくても同じサイトのデータならばロボットによって取得するようにする。

【0093】例えば、以下のような番号を名前とするようなURLは一時的にのみ存在している可能性が高い。

http://www.tsb.co.jp/foo/1246389.html

このような場合には、このファイルを再び取得するのではなく、このファイルへのリンクを張っているファイルを取得し、そのファイルからリンクを辿った先のファイルを取得する。

【0094】図11の検索装置では、プリフェッチしたものが将来使われると仮定している。ここでプリフェッチする対象は、文字情報、画像、音声、動画などのメディアを任意に選択できるものとする。例えば、記憶容量の制約から文字情報のみをプリフェッチすることに指定したが、そのページに動画が入っていた場合には、その動画は利用者が参照したときにネットワークを介して取りに行くか、表示されないかのいずれかになる。

【0095】次に、図10や図11の検索装置におけるページの取得頻度に関して説明する。ロボットは、同じURLのページを定期的に取得しに行くが、その際、対象ページの更新頻度に応じた頻度で該ページを取り直すのが好ましい。すなわち、対象ページが統計的に一日に変更される回数に比例した回数だけ、該ページを取得しに行く。ただし、指定したデータが消滅したならば、二度とそのデータを取りに行かないようにする。また、取

得したデータがハイパーリンクとなっている場合には、リンク先の情報も取りに行くことも可能である。

【0096】また、指定したサイト群やURL群のデータについては、利用者がリロード要求を出しても、それに応じないようにする。これにより、検索サーバから同じURLに対する一定回数以上の要求がでないことが保証される。

【0097】次に、図10や図11の検索装置における検索対象に関して説明する。本実施形態では、ロボットで収集したデータもプロキシーのキャッシュの中に入れておき、利用者が直接要求したデータと同じ場所で管理する。

【0098】ここで、参照したコンテンツが暗号化されていない有料データのこともあるし、利用者のプライバシーの問題もあるので、検索システムが検索対象とするデータに削除が加えられるようにしても良い。

【0099】削除の与え方としては、以下の条件を1つ以上組み合わせたものとする。

（1）ロボットで収集したものに限る、（2）プロキシーとしてデータを保持しているものに限る、（3）同じ名前もしくはアドレスを持つ情報については最新のものだけに限る、（4）CGIなどにより動的もしくは対話的に生成された情報は除く、（5）指定したサイト群やURL群に限る。

【0100】例えば、図10において、データをキャッシュ604に入れるとき、そのデータの取得状況も記録しておく。すなわち、そのデータが、ロボットで収集したものか、利用者が直接要求したものか、CGIなどにより動的もしくは対話的に生成されたものか（これはURLのパス名にCGIやBINという文字を含むかどうかで判定する）、指定されたサイト群かURL群かなどの情報も、データと一緒に記録しておく。そして、管理者がどの種類のデータはキャッシュ内のデータについて検索が可能かどうかを指定できるようにしておく。検索システムでは、この指定に従って、条件の合うものだけをデータベース化する。

【0101】次に、図10や図11の検索装置における収集データのアドレスの付け替えについて説明する。本実施形態では、収集したデータを検索装置のキャッシュに格納する際に、該収集データのアドレスもしくはURLを付け変えて格納しておいても良い。すなわち、データの位置がネットワークのある場所から検索装置内のキャッシュに移動したのであるから、ドメイン名を検索装置のドメイン名に変えるようにする。次に、パス名の先頭に元のドメイン名を付加する。例えば、以下のようにする。

【0102】元のURL　http://www.foo.co.jp/bar/index.html
検索装置のドメイン名　www.search.co.jp
新たなURL　http://www.search.co.jp/www.foo.co.jp/bar/index.html
このようにすることにより、データのミラー化が実現できる。

【0103】次に、図10や図11の検索装置における収集データの時刻管理について説明する。本実施形態では、収集データに更新時刻データも付与して管理するようにしても良い。通常のプロキシーのように同じアドレス（URL）に対しては、最新のデータのみを保持するだけでなく、過去のデータも管理して保持する。ここでの時刻は、そのデータが有効になった時刻、あるいはそれに加えて無効になった時刻とを持つ。

【0104】有効になった時刻は、同一URLで内容が更新されたような場合には、サーバーから通知される更新時刻が変化するので、その時刻が無効になった時刻になり、データそのものが消滅した場合には、アクセスに行ったことにより消滅したことが判った時刻とする。

【0105】アドレス（URL名）は、時刻管理をするために付け替えて管理する。まず、データの位置がネットワークのある場所から検索装置内のキャッシュに移動したのであるから、ドメイン名を検索装置のドメイン名に変える。次に、パス名の先頭に元のドメイン名を付加する。例えば、以下のようにする。

【0106】元のURL　http://www.foo.co.jp/bar/index.html
検索装置のドメイン名　www.search.co.jp
新たなURL　http://www.search.co.jp/bar/index.html
さらに、これに時刻の情報も付与する。例えば、1996年3月23日16：39から1996年4月30日10：23まで有効であったデータならば、以下のようにする。

【0107】http://www.search.co.jp/www.foo.co.jp/bar/index.html/19960323163 9－19960430 1023
また、以下のような変形も考えられる。

【0108】http://www.search.co.jp/www.foo.co.jp/bar/index.html/1996.3.23.16.39－1996.4.30.10.23
なお、以上説明したさた本発明の実施の形態における各構成は、相当する手順あるいは手段をコンピュータに実行させるためのプログラムを作成し、これをコンピュータに実行させることにより実現可能である。

【0109】また、上記プログラムを機械読取り可能な媒体に記録し、コンピュータがこの媒体からプログラム

を読取って実行するように構成することも可能である。本発明は、上述した実施の形態に限定されるものではなく、その技術的範囲において種々変形して実施することができる。

【0110】

【発明の効果】本発明によれば、データの更新頻度に応じて異なったデータベースにてデータを管理することができる。この結果、例えば、そのデータベースが管理するデータの更新頻度の高さに応じて計算機等の持つ処理能力を設定することができ、ネットワーク上に分散された膨大なデータを効果的に管理することができる。

【0111】また、本発明によれば、検索システムにプロキシー機能をも内蔵させたので、プロキシーに格納されているデータを検索し提示することができる。この結果、例えば、極めて更新頻度が高いデータに対しても、検索サービス・参照サービスを提供することができる。

【図面の簡単な説明】

【図1】本発明の第1の実施形態に係る検索装置の構成例を示す図

【図2】同検索装置の処理手順の一例を示すフローチャート

【図3】最終更新時刻によってデータベースを分担する方法を説明するための図

【図4】同実施形態に係る検索装置の他の構成例を示す図

【図5】同検索装置の処理手順の一例を示すフローチャート

【図6】同実施形態に係る検索装置のさらに他の構成例を示す図

【図7】同検索装置の処理手順の一例を示すフローチャート

【図8】同実施形態に係る検索装置のさらに他の構成例を示す図

【図9】複数のネットワークが接続された場合のシステム構成の一例を示す図

【図10】本発明の第2の実施形態に係る検索装置の構成例を示す図

【図11】本発明の第2の実施形態に係る他の検索装置の構成例を示す図

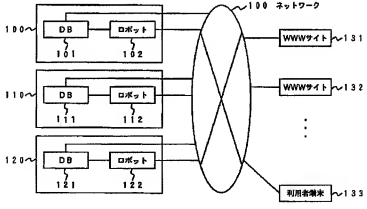
【図12】同検索装置の処理手順の一例を示すフローチャート

【図13】頻繁に変化するページのサンプリングを説明するため図

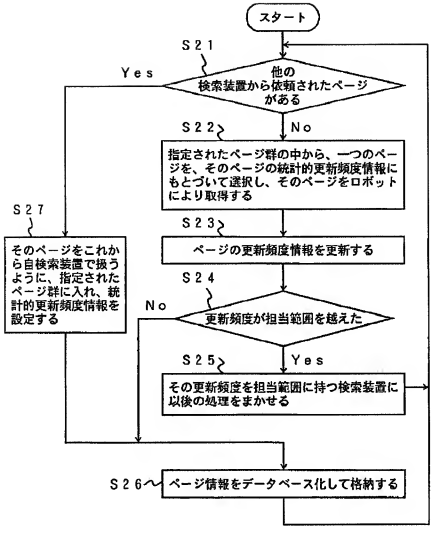
【符号の説明】

100，500～504，600…ネットワーク
100，110，120，200，300，401，410，601…検索装置
102，112，122，204，602…ロボット
101，101－1，101－2，111，121，605…データベース
131，132…WWWサイト
133…利用者端末
301，301－1，301－2，607…データベース・フロントエンド（DBF）　603…キャッシュ
604…データベース化部
606…WWWフロントエンド
708…ユーザ要求記録部

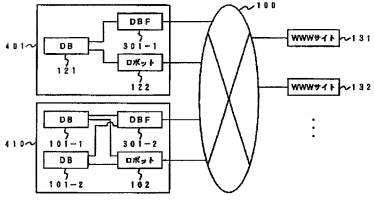
【図1】



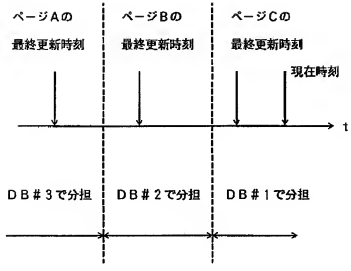
【図2】


スタート
S21
他の検索装置から依頼されたページがある
Yes
No
S22
指定されたページ群の中から、一つのページを、そのページの統計的更新頻度情報にもとづいて選択し、そのページをロボットにより取得する
S23
ページの更新頻度情報を更新する
S24
更新頻度が担当範囲を越えた
No
Yes
S25
その更新頻度を担当範囲に持つ検索装置に以後の処理をまかせる
S27
そのページをこれから自検索装置で扱うように、指定されたページ群に入れ、統計的更新頻度情報を設定する
S26
ページ情報をデータベース化して格納する

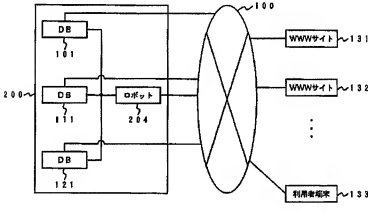
【図8】


100
401
DB
121
DBF
301-1
Dボット
122
WWWサイト
131
WWWサイト
132
410
DB
101-1
DB
101-2
DBF
301-2
Dボット
102

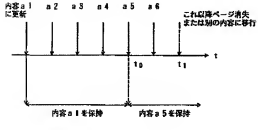
【図3】


ページAの
最終更新時刻
ページBの
最終更新時刻
ページCの
最終更新時刻
現在時刻
t
DB＃3で分担
DB＃2で分担
DB＃1で分担

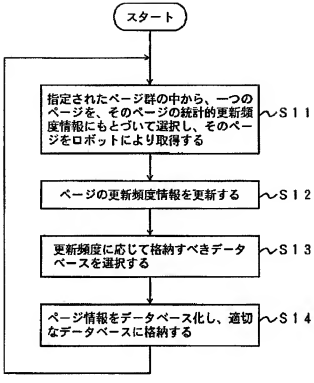
【図4】


100
DB
101
WWWサイト
131
200
DB
111
ロボット
204
WWWサイト
132
DB
121
利用者端末
133

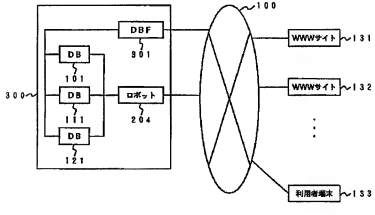
【図13】


内容a1
に更新
a2
a3
a4
a5
a6
これ以降ページ消失
または別の内容に移行
t
t0
t1
内容a1を保持
内容a5を保持

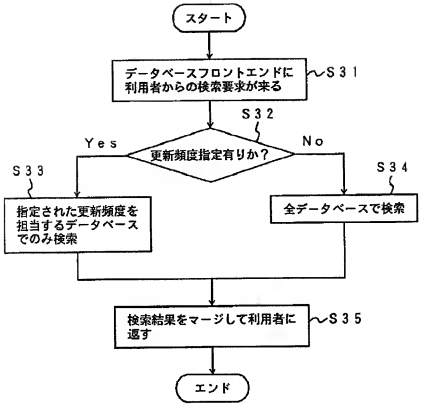
【図5】


スタート
指定されたページ群の中から、一つのページを、そのページの統計的更新頻度情報にもとづいて選択し、そのページをロボットにより取得する
S11
ページの更新頻度情報を更新する
S12
更新頻度に応じて格納すべきデータベースを選択する
S13
ページ情報をデータベース化し、適切なデータベースに格納する
S14

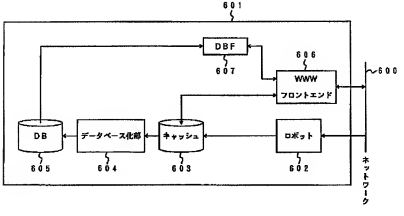
【図6】


100
300
DBF
301
DB
101
DB
111
DB
121
ロボット
204
WWWサイト
131
WWWサイト
132
利用者端末
133

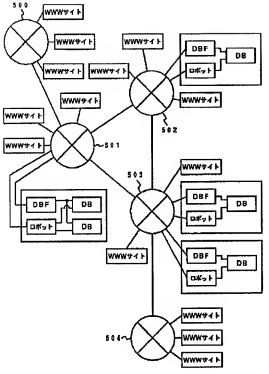
【図7】


スタート
データベースフロントエンドに利用者からの検索要求が来る
S31
S32
更新頻度指定有りか？
Yes
No
S33
S34
指定された更新頻度を担当するデータベースでのみ検索
全データベースで検索
検索結果をマージして利用者に返す
S35
エンド

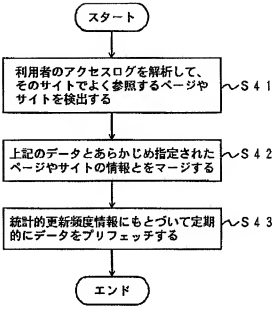
【図10】


601
DBF
607
606
WWW
フロントエンド
600
DB
605
データベース化部
604
キャッシュ
603
ロボット
602
ネットワーク

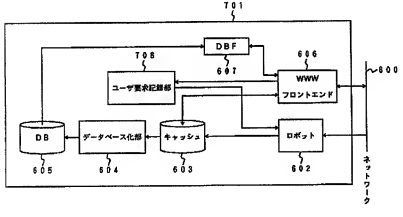
【図9】



【図12】



【図11】



フロントページの続き

(72)発明者　山根　徹也
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内